ED 380 497 TM 022 860

AUTHOR	Donoghue, John R.
TITLE	A Preliminary Study of the Effects of Within-Group Covariance Structure on Recovery in Cluster Analysis. Research Report RR-94-46.
INSTITUTION	Educational Testing Service, Princeton, N.J.
PUB DATE	Sep 94
NOTE	55p.; Version of a paper presented at the Annual Meeting of the American Educational Research Association (Atlanta, GA, April 12-16, 1993).
PUB TYPE	Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	Algorithms; *Analysis of Covariance; *Cluster Analysis; *Correlation; *Group Membership; Monte Carlo Methods; Statistical Studies
IDENTIFIERS	Covariance Structure Models; Hierarchical Models; Variance (Statistical); *Within Group Differences

ABSTRACT

Monte Carlo studies investigated effects of within-group covariance structure on subgroup recovery by several widely used hierarchical clustering methods. In Study 1, subgroup size, within-group correlation, within-group variance, and distance between subgroup centroids were manipulated. All clustering methods were strongly affected by within-group correlation; negative correlation yielded much poorer recovery. Smaller effects were found for the interaction of clustering method and within-group variance. Study 2 separated effects of direction of correlation from the direction of differences in the subgroup centroids. Subgroup size, within-group correlation, direction of the vector separating subgroup centroids, and distance between subgroup centroids were manipulated. Superior recovery was associated with within-group correlation that matched the direction of subgroup separative. Overall, the EML algorithm of the Statistical Analysis System yielded best recovery, followed closely by Ward's method, average linkage, and a version of the beta-flexible algorithm. Several alternative measures are discussed. Six tables and eight figures present analysis data. (Contains 52 references.) (Author/SLD)

RR-94-46

RESEARCH REPORT

A PRELIMINARY STUDY OF THE EFFECTS OF WITHIN-GROUP COVARIANCE STRUCTURE ON RECOVERY IN CLUSTER ANALYSIS

John R. Donoghue

Educational Testing Service
Princeton, New Jersey
September 1994

A Preliminary Study of the Effects of Within-group Covariance Structure on Recovery in Cluster Analysis

John R. Donoghue

Educational Testing Service

This research was supported by Educational Testing Service Research Allocation Project 794-10. An earlier version of this paper was presented at the Educational Statisticians SIG at the annual meeting of the American Educational Research Association in Atlanta, GA, April 1993. The author would like to thank Carolyn Crone and Charles Lewis for their numerous helpful comments.

Abstract

Two Monte Carlo studies investigated the effects of within-group covariance structure on subgroup recovery by several widely used hierarchical clustering methods. Data sets were 100 bivariate observations from two subgroups, generated according to a finite normal mixture model. In Study 1, subgroup size, within-group correlation, within-group variance, and distance between subgroup centroids were manipulated. All clustering methods were strongly affected by within-group correlation; negative correlation yielded much poorer recovery. Smaller effects were found for the interaction of clustering method with within-group variance.

Study 2 separated the effects of direction of correlation from the direction of differences in the subgroup centroids. Subgroup size, within-group correlation, direction of the vector separating subgroup centroids, and distance between subgroup centroids were manipulated. Superior recovery was associated with within-group correlation that matched the direction of subgroup separation. Overall, the EML algorithm of SAS yielded best recovery, followed closely by Ward's method, average linkage, and a version of the beta-flexible algorithm, although several interactions were noted.

The results are interpreted according to the weakness of the (squared) Euclidean distance as a measure of (dis)similarity for cluster analysis. Several alternative measures are discussed, and promising alternatives are identified for future investigation.

Cluster analysis is typically used to try to identify relatively homogeneous subgroups within a more heterogeneous population. For example, in one application of cluster analysis, several investigators have suggested that the umbrella term "learning disabilities" actually encompasses a variety of disorders, with different etiologies and strategies for treatment. Cluster analysis has been used to try to identify putative subtypes of learning disability. In such applications, the researcher is confronted by a bewildering variety of clustering methods, each with a number of options. Indeed, Milligan and Cooper (1987, pp. 349-350) describe cluster analysis as a "complex series of tasks that must be carefully executed to obtain a proper clustering of the user's data."

Numerous studies have been conducted to examine various aspects of clustering. Features examined include comparison of clustering algorithms (e.g., Belbin, Faith, & Milligan, 1992; Blashfield, 1976; Milligan, 1979, 1989a; Scheibler, & Schneider, 1985); the effect of various types of "error" in the data to be clustered (Milligan, 1980); variable standardization (Milligan, & Cooper, 1988); differential variable weighting (De Soete, 1986, 1988; Milligan, 1989b; Donoghue, 1994ab); procedures to determine the number of clusters (e.g., Milligan, & Cooper, 1985); and procedures to compare clustering solutions (Milligan, & Cooper, 1986).

Although these studies have obtained a variety of useful findings, one aspect of the clustering process has not been systematically examined: the within-group covariance structure. Three general approaches have been taken in the past. In the first group of studies, exemplified by the work of Milligan and colleagues (e.g., Belbin, Faith, & Milligan, 1992; Milligan, 1980, 1981, 1985, 1989a; Milligan, & Cooper, 1985, 1986, 1988), variables are generated to be independent within subgroups. Within-group variances are usually independently sampled from some univariate distribution. The second group of studies, such as Blashfield (1976) and Blashfield and Morey (1980), generate data from mixtures of multivariate normal distributions with "complex covariance structures." The structures are typically idiosyncratic and often vary in unspecified ways from data set to data set. A variation of this approach is to generate correlation matrices randomly from the population of such matrices with a given eigenstructure (Breckenridge, 1989). A third, related alternative is the use of specific empirical data sets, such as Fisher's (1936) famous data consisting of measurements of *Iris* plants. This approach was adopted by Mezzich and Solomon (1980) and Dreger, Fuller, and Lemoine (1988).

There seems to be good reason to expect the within-group covariance structure to affect the behavior of clustering algorithms. Prior to clustering, hierarchical clustering methods convert two-mode (variables by entities) multivariate data into a single-mode (entities by entities) univariate measure of similarity.[1] The vast majority of clustering studies (applied and simulation) use d_E, Euclidean distance (or the squared distance) between entities *i* and *j* as similarity measure:

$$d_E(ij) = \sqrt{\sum_{p=1}^{P} (x_{ip} - x_{jp})^2} \quad , \qquad (1)$$

where P is the number of variables used in the clustering. Basic geometry of vector spaces reveals that Euclidean distance is correct only if computed on an orthogonal basis. If the variables are correlated, the Euclidean distance can misrepresent the distance between two points. Thus, within-group covariance may adversely affect the ability of various clustering methods to recover subgroup structure. Note that most hierarchical clustering algorithms can make use of any one-mode measure of similarity. However, d_E is widely used and is the default measure for most clustering software. Thus, it is the only measure of similarity that will be examined in this paper.

The two studies reported in this paper only examined the limited case of two subgroups in

[1] The general term "measure of similarity" will be used in this paper to denote both true measures of similarity, such as correlations and coefficients of concordance, and measures of *dis*similarity, such as distances.

two dimensions, but this case was examined in some detail. A subsequent paper will examine the degree to which these findings extend to more subgroups and higher dimensional clustering problems. The studies sought to address the following questions:

1) Does the within-group covariance structure affect the ability of commonly used clustering methods to recover known subgroup structure? If so, which aspects of that structure seem most important?

2) Do clustering methods differ in their sensitivity to within-group covariance structure?

3) Overall, which clustering methods work best for this type of data?

Study 1

Method

Study 1 used Monte Carlo methods to investigate systematically the effects of various within-group covariance structures on the ability of clustering algorithms to recover a known subgroup structure.

Design

The chief variable of interest was the structure of the within-group covariance matrices. Three aspects of the within-group covariance matrices were manipulated.

1) R1: The within-group correlation of subgroup 1 (7 levels)-- r_1 = -.9, -.7, -.3, 0.0, .3, .7, .9.

2) R2: The within-group correlation of subgroup 2 (7 levels)-- r_2 = -.9, -.7, -.3, 0.0, .3, .7, .9.

3) VAR: The ratio of the of within-group variances. Within a subgroup, the variances of each variable were equal ($\sigma^2_{1g} = \sigma^2_{2g}$). VAR was manipulated by setting both of the variances of the second subgroup to a *k* constant times the variance within the first subgroup ($\sigma^2_{12} = k\sigma^2_{11}$, $\sigma^2_{22} = k\sigma^2_{21}$). The ratio VAR had 5 levels-- 1:9; 1:4, 1:1, 4:1, and 9:1.

In generating the data, two other "nuisance variables" were manipulated, because of their consistently large effects in other studies:

4) PROB: Probability of each of the subgroups in the population (p_1 and p_2) was manipulated by varying the sizes ($n_1 = N^*p_1$, $n_2 = N^*p_2$) of the subgroups (2 levels)--Equal-sized subgroups (p_1=.50, p_2=.50) or unequal subgroup sizes (p_1=.90, p_2 =.10).

5) DIST: Separation of subgroups. This was defined in terms of D_M, the Mahalanobis distance between the population centroid vectors of the subgroups:

$$D_M(12) = \sqrt{(\mu_1-\mu_2)'\Sigma^{-1}(\mu_1-\mu_2)} \qquad (2)$$

This distance was based on the pooled within-group covariance matrix: $\Sigma = p_1{}^*\Sigma_1 + p_2{}^*\Sigma_2$. DIST had 3 levels--$D_M$ = 2, D_M = 4, or D_M = 6.

Finally, each data set was analyzed using a variety of hierarchical clustering algorithms:

6) METHOD: The method of cluster analysis (10 levels)--which are discussed below.

The first five factors were fully crossed to yield 1470 conditions. Twenty data sets were generated for each condition, according to the procedure given below. Each data set was then analyzed by each of the 10 hierarchical clustering methods, yielding a total of 294,000 cluster analyses. For each cluster analysis, the solution for the correct number of subgroups (i.e., G=2) was used as the result for that clustering method.

Data Generation

Data were generated using the DATA step in SAS. Clusters were generated according to a model of a finite mixture of normal distributions (e.g., Titterington, Smith, & Makov, 1985; McLachlan, & Basford, 1988). All data sets consisted of 100 observations. The centroids were separated by both variables equally (i.e., $\mu_2 - \mu_1 = k \cdot 1, k > 0$). To ensure that the data generation procedures worked properly, the Mahalanobis distance between the two known subgroups was computed for each data set. Table 1 presents descriptive statistics for D_M.

Insert Table 1 about here

The means for each of the conditions are very close to the desired values, supporting the validity of the data generation procedures.

Cluster Algorithms

Each data set was analyzed 10 times, corresponding to different hierarchical clustering algorithms and measures of similarity. SAS PROC CLUSTER (SAS Institute, 1988) was used. The clustering methods were:

1) Single linkage, Euclidean distance;

2) Complete linkage, Euclidean distance;

3) Average linkage, Euclidean distance;

4) Average linkage, squared Euclidean distance;

5) Ward's (1963) method (minimum variance), Euclidean distance;[2]

6) Ward's method (minimum variance), squared Euclidean distance;

7) EML-SAS's maximum likelihood hierarchical clustering procedure, which is a modification of Ward's method to alleviate the method's tendency to yield equal-sized clusters;

8) Lance and Williams' (1967) beta-flexible linkage ($\beta = -.25$), Euclidean distance;

9) Beta-flexible linkage ($\beta = -.50$), Euclidean distance; and

10) Beta-flexible linkage ($\beta = -.75$), Euclidean distance.

[2] This is not a true application of Ward's method, which requires squared distances to minimize within-group variance. However, because this option is widely available in clustering software, it was decided to include the method in the study.

Note that single linkage yields identical results for distance and squared distance measures of similarity. This is also true of the complete linkage method. The SAS implementation of the beta-flexible method only allows the use of distances. Thus, only the distance measure was used with these three methods. The clustering methods were chosen because: a) they are widely used (single linkage, complete linkage, average linkage, and Ward's method); b) they have performed well in previous studies (average linkage, Ward's method, and beta-flexible); or c) they are designed to rectify a known weakness of another algorithm (EML). The values of β used for the beta-flexible clustering method are based on the study by Milligan (1989a). For a discussion of these algorithms, the reader is referred to standard introductions to cluster analysis (e.g., Everitt, 1974; Lorr, 1983); Milligan (1989a) and Belbin, Faith, and Milligan (1992) contain discussions of the beta-flexible methods, and the SAS documentation (SAS Institute, 1988) is the primary reference for the EML algorithm.

Outcome Measure

The outcome measure for the study was the Hubert and Arabie (1985) modification of Rand's (1971) statistic, which will be denoted HA-Rand. The index was computed between each cluster solution and the true subgroup membership used to generate the data. This index is based on examining pairs of subjects, and determining whether they are classified into the same or different subgroups. A value of zero reflects chance agreement with the true membership, and 1.0 reflects perfect agreement. A study by Milligan and Cooper (1986) supports the accuracy of Hubert and Arabie's modification.

Analyses

To summarize the results, a full factorial analysis of variance was conducted, as in Milligan (1980, 1981, 1989a). The HA-Rand index served as the dependent variable. The independent variables were the design factors used to generate the data and the cluster algorithm used to analyze the data. This analysis was also performed on the variance stabilizing arc-sine transformation of the HA-Rand index. The results were very similar for both analyses. Therefore, only the results in the original metric will be discussed here.

Given the very large amount of data (294,000 cluster analyses), it was not surprising that all of the main effects and interactions were highly significant ($p < .0001$). The purpose of the ANOVA was to summarize the data and help to highlight the more important effects. Therefore, a measure of effect size was adopted in place of traditional significance testing. Usually, η^2 would be used in this context:

$$\eta^2 = \frac{SS_{effect}}{SS_{Tot}} \quad . \tag{3}$$

However, this index has a disadvantage in large designs, namely that the denominator contains not only error variance and systematic variance of interest, but also irrelevant systematic variance of other factors in the design. The larger the design becomes, the more apparent this effect becomes. In the present case, this defect is particularly noisome because one of the factors, distance between the subgroup centroids, has a very large effect and so serves to obscure the effects of other factors. Therefore, an alternate version, η^2_{alt} (Tabachnick, & Fidell, 1983, p. 47), was used:

$$\eta^2_{alt} = \frac{SS_{effect}}{SS_{effect} + SS_{error}} \quad . \tag{4}$$

Note that, unlike the other formulation, this version, η^2_{alt}, does *not* sum to 1.0. A practical criterion of $\eta^2_{alt} > .03$ was selected. Although this criterion is somewhat arbitrary, it was chosen for two reasons. First, Monte Carlo studies in other domains have used similar criteria (e.g., $\omega^2 \geq 3\%$ was used by Anderson, & Gerbing, 1984, and Gerbing, & Anderson, 1985). Second, preliminary examination of the results revealed that this criterion functioned well in separating effects of some size from effects associated with only minor differences in the means.

Results of Study 1

The salient effects identified in the ANOVA are listed in Table 2. DIST, the distance between the subgroup centroids, had a particularly large effect; as the D_M increases and the clustering problem becomes easier, cluster recovery improves. All the main effects are salient with the notable exception of VAR. In addition, several of the two-way interactions and one three-way interaction were flagged as salient. Overall, these effects account for more than 85% of the systematic variance in the data.

Insert Table 2 about here

Within-group Covariances

Because the focus of the study was on within-group covariance matrices, we begin by examining the effects of within-group correlation. These effects are modified by several salient two-way interactions. Figure 1 gives a response surface for the cell means of the R1 by R2 interaction. The x and y axes are R1 and R2, and the height of the surface represents the average value of the HA-Rand index. In general, recovery of subgroup structure is better when the variables are positively correlated. Correlations of .7 or greater lead to the highest recovery values, although gains beyond the point where both of the correlations are 0 are not large. In contrast, when one or both of the subgroups have negative correlation, recovery is much worse. For the condition R1 = R2 = -.9, recovery is only slightly better than chance. The larger effect connected with R1 (the correlation in subgroup 1) is an artifact of the design; in half of the data sets, subgroup 1 comprised 90% of the sample.

Insert Figure 1 about here

Tables 3 and 4 give the means for the VAR by R1 and VAR by R2 interactions. In both cases, there is a strong effect of within-group correlation when it is associated with larger variances. When the ratio becomes 1:1, within-group correlation has a smaller effect, and as the variance of the other subgroup becomes larger, the effect of the correlation is relatively minor.

Insert Tables 3 and 4 about here

Combining the results of these three interactions, it appears that the pooled within-group covariance matrix Σ_w has an effect on the recovery of subgroup structure; positive covariance in Σ_w leads to somewhat better recovery, while negative covariance leads to worse recovery.

Clustering Methods

Although the main effect of clustering method was identified as salient, it must be qualified by considering the Method by DIST, Method by PROB, and Method by VAR interactions. Figure 2 shows the interaction of clustering method with PROB. Average linkage is insensitive to relative subgroup size. The EML and complete linkage methods yield similar results for equal-sized subgroups and unequal-sized subgroups. Ward's method and the beta-flexible clustering method yield better recovery than EML for equal-sized subgroups, and worse than EML for unequal-sized subgroups.

Insert Figure 2 about here

The interaction of Method with DIST is portrayed in Figure 3. Although the size of differences between methods varies, there is little change in the relative ordering. A few exceptions occur for $D_M = 2$, where more negative values of β produce slightly superior results to $\beta = -.25$ for the beta-flexible clustering method. Also, for $D_M = 2$, complete linkage clustering produced recovery superior to average linkage. In general, however, methods are ordered similarly for each of the values of D_M.

Insert Figure 3 about here

The third salient interaction was Method by VAR. Figure 4 gives cell means for the extreme 1:9 and 9:1 variance ratios, and the middle value of 1:1; the omitted values generally fell between the 1:1 line and the appropriate extreme condition. One aspect of Figure 4 is striking; the average linkage clustering method is sensitive to the variances, and yields best recovery with homogeneous subgroup variances. Indeed, for homogeneous variances average linkage becomes one of the better clustering methods. EML is less sensitive to variances, although conditions for which subgroup 1 (the larger subgroup in the unequal probability of membership condition) had a larger within-group variance yielded somewhat lower recovery. The beta-flexible method yielded better recovery when subgroup 1 had a smaller variance. For $\beta = -.25$, the method yielded similar recovery when subgroup 1 had variance which was equal to or larger than that of subgroup 2. Ward's method using squared distance showed a similar pattern of results to that of the beta-flexible method with $\beta = -.25$, although the effect was not as large.

Insert Figure 4 about here

Other Effects

The three-way interaction between DIST, PROB, and R1 was identified as salient, as were two of its component two-way interactions, DIST by R1 and PROB by R1. Figure 5 portrays the three-way interaction. Increasingly negative values of R1 have the greatest effect when the distance between the subgroup centroids is large and when subgroup 1 comprises the bulk of population. The effects of the two-way interactions are also evident from the plot. The deleterious effect of a negative R1 on recovery increases as the subgroup centroids become farther apart. It appears that, for subgroups which are closer together, subgroup overlap masks the effect of a negative within-group correlation. Also, the effect of R1 is larger when subgroup 1 is 90% of the population than when it is 50%. This provides further support for the speculation that the structure of the pooled within-group covariance matrix affects recovery.

Insert Figure 5 about here

Discussion of Study 1

The resulic of Study 1 are unequivocal; within-group covariance structure does influence the ability of all cluster analysis methods to recover known subgroups. The within-group correlation was found to have a strong effect. Negative within-group correlation was associated with inferior cluster recovery. The effect of increasing positive ccrrelation was much less dramatic, and only minimal gains were found as the correlation increased from 0.3 to 0.9. There is some evidence that the effect of correlation may be driven by the pooled within-group covariance matrix, Σ_w.

The relative sizes of the within-group variances interacted with cluster methods. The average linkage method performed best with homogeneous variances. On the other hand, Ward's method , EML, and the beta-flexible method all yielded slightly better recovery when subgroup 1 (the larger subgroup for the unequal probability condition) had a smaller variance than subgroup 2.

One aspect of these findings is problematic, however. The relative position of the subgroup centroids did not vary; the line between the subgroup centroids always fell at a 45° angle from the x-axis. Thus, it is not clear whether superior recovery is associated with a positive within-group correlation *per se*, or whether superior recovery is associated with a within-group correlation which matches the direction of subgroup separation. To determine the answer to this question, a second, smaller follow-up study was performed.

Study 2

Study 2 sought to separate the effects of direction of correlation from the direction of differences in the subgroup centroids.

Method

The design of Study 2 was very similar to that of Study 1. Data sets again consisted of 100 bivariate observations drawn from a two-group normal mixture model. The following features of the data were varied:

1) R1: Within-group correlation of subgroup 1 (3 levels)--r_1 = -.7, 0.0, .7;

2) R2: Within-group correlation of subgroup 2 (3 levels)--r_2 = -.7, 0.0, .7;

3) DIST: Separation of subgroups defined as in Study 1 (3 levels)--D_M = 2, 4, or 6;

4) PROB: Probability of the subgroups in the population, defined as in Study 1 (2 levels)--Equal-sized subgroups or unequal subgroup sizes;

5) ANGLE: Direction of separation of subgroup centroids (8 levels)--The direction of separation between subgroup centroids was defined as the angle θ between a line drawn through the centroids and the x-axis. The centroid for subgroup 1 was always located at (0,0). The centroid for subgroup 2 was then determined to create the desired angle between the x-axis and the line connecting the centroids. The eight levels of θ were: 0°(falling on the

x-axis), 30°, 45°, 60°, 90° (falling on the y-axis), 120°, 135°, and 150°. The means were determined from θ: $\mu_{x2} = k \cos(\theta)$ and $\mu_{y2} = k \sin(\theta)$, where k was determined by the level of DIST.

The ratio of the of within-group variances (VAR from Study 1) was held constant at 1:1. Each data set was analyzed using a variety of clustering algorithms:

6) METHOD: Method of cluster analysis (10 levels)--The same methods that were used in Study 1.

The first five factors were fully crossed to yield 432 conditions. Ten data sets were generated for each condition, using the same procedures as in Study 1. Each data set was then analyzed according to the 10 hierarchical clustering algorithms, yielding a total of 43,200 cluster analyses. The HA-Rand index was again used as the outcome measure.

Results of Study 2

To summarize the results, a full factorial ANOVA was again conducted. The HA-Rand index served as the dependent variable. The independent variables were the design factors used to generate the data and the cluster algorithm used to analyze the data. Again, the practical criterion of effect size measure $\eta^2_{alt} > .03$ was adopted in place of traditional significance testing.

The chief variables of interest in this analysis are R1, ANGLE, and the ANGLE by R1 interaction. Two patterns of outcomes are of particular interest.

I) If the results of Study 1 were due to positive correlation *per se*, then we expect a strong main effect for R1, with improved recovery associated with positive correlations. If this be the case, we have no reason to expect any particular effect for ANGLE, nor do we expect a large ANGLE by R1 interaction.

II) If the results of Study 1 were due to concordance between the direction of the vector separating the subgroup centroids and the correlation, then we expect a large ANGLE by R1 interaction, with the best recovery occurring when: a) $\theta=45°$ and r=.7; and b) $\theta=135°$ and r=-.7. Conversely, we expect worst recovery when: a) $\theta=45°$ and r=-.7; and b) $\theta=135°$ and r=.7. The main effects of ANGLE and R1 are of less interest under this scenario.

The salient effects identified in the ANOVA are listed in Table 5. These effects account for approximately 88% of the systematic variance observed in the HA-Rand index.

Insert Table 5 about here

The three-way interaction between ANGLE, PROB, and R1 was identified as salient, as were two of its component two-way interactions, ANGLE by R1 and PROB by R1. Figure 6 portrays the three-way interaction. As predicted under scenario II, the best recovery appears for: a) $\theta=45°$ and r=.7, and b) $\theta=135°$ and r=-.7. We expected the worst recovery when: a) $\theta=45°$ and r=-.7, and b) $\theta=135°$ and r=.7. Recovery was indeed poor under these conditions, but nearby angles produced slightly worse recovery ($\theta=30°$ for r=-.7 and $\theta=120°$ and $\theta=150°$ for r=.7). Nonetheless, the basic pattern of results predicted by scenario II was found. The effect PROB is to moderate the above interaction; the pattern of means is identical. For equal probabilities of subgroup membership, the interaction is much smaller than it is for unequal probabilities. This result is an expected artifact of the design. When the probabilities are unequal, R1 is associated with the larger subgroup, and thus has a larger effect under those circumstances. This observation also provides further support for the speculation that the structure of the pooled within-group covariance matrix affects recovery.

Insert Figure 6 about here

Table 5 reveals that three additional two-way interactions were also identified as salient: R1 by R2, Method by PROB, and Method by DIST. Examination of the means (not shown) indicated that the R1 by R2 and Method by PROB interactions replicated the effects of Study 1, which were portrayed in Figures 1 and 2. The remaining interaction, Method by DIST, differed somewhat from the effect found in Study 1. In Study 1, the clustering methods were ordered similarly for each level of DIST, although the size of the effect of DIST differed fcr various methods. In Study 2, there *were* changes in the ordering of methods. When $D_M = 2$, average linkage yielded lower recovery than every method except single linkage, and Ward's method using squared distances yielded the highest recovery. When $D_M = 4$ and $D_M = 6$, average linkage is second only to EML. Recall that Study 1 found a significant interaction between the ratio of subgroup variances and clustering methods, with average linkage performing very well when the subgroup variances were equal. The good results for average linkage found in Study 2 support that finding.

Discussion of Study 2

Before considering the results of Study 2 in detail, we consider the possibility that some of the findings obtained may be an artifact of the methods used to generate the data. The Mahalanobis distance between the subgroup centroids is a function of the within-group covariance; the larger the within-group correlation, the farther apart (in a Euclidean sense) the centroids. Figure 7 plots the centroids for each value of ANGLE in Study 2, when $\Sigma_1 = \Sigma_2$. The centroid of subgroup 1 is located at the origin. The points which are connected by lines represent the location of the centroid of subgroup 2 for various values of within-group correlation. The D_M between the centroids is 4 in every case.

Insert Figure 7 about here

The Euclidean distances from the origin to each of the centroids closely follows the pattern of results of Figure 6, raising the possibility that these results may be an artifact.

This comparison of Figure 7 with Figure 6 relies on the false assumption that the Euclidean distances between the subgroup centroids accurately reflect the inherent difficulty of the clustering task. Define the "inherent difficulty of clustering" as the theoretical probability of misclassification (PMC), i.e, the probability, under an optimal decision rule, that an entity from subgroup 1 will be falsely assigned to cluster 2 + the probability that an entity in subgroup 2 will be falsely assigned to cluster 1. For simplicity, consider the subset of Study 2 in which $p_1 = p_2$ and $\Sigma_1 = \Sigma_2$. Under these conditions, a result from discriminant analysis (e.g., Siotani, 1982) is that the theoretical PMC *depends only upon the Mahalanobis distance between the centroids* and is given by:

$$PMC = 2 \cdot \Phi\left(-\frac{D_M}{2}\right) \quad ,$$

where $\Phi(\cdot)$ is the (cumulative) distribution function of the normal distribution. D_M is identical for each of the conditions depicted in Figure 7, and so the inherent dif"iculty of the clustering problem is identical.

To illustrate this point, the portion of the data from Study 2 which met the conditions listed above ($p_1 = p_2$ and $\Sigma_1 = \Sigma_2$) was reanalyzed. Next, parametric normal theory discriminant function analyses were conducted for each data set, using SAS's PROC DISCRIM (SAS Institute, 1988). Resubstitution allocation assigned each observation to one of the clusters. The HA-Rand index was computed for the resulting allocation. These results are optimal, in the sense that they make use of the informatio~~ about subgroup membership; the overall HA-Rand index mean of .80 for the discriminant analysis results is much higher than is the mean of .56 for the clustering results. Even the best of the clustering algorithms for these data, Ward's method, only yielded a mean of .66.

An ANOVA of the clustering results (not shown) indicated several salient effects. The most noteworthy were the main effect of within-group correlation (R) and the large interaction between R and ANGLE. Examination of the means (not shown) revealed that the effects followed the same pattern as that for the analysis of the full data. On the other hand, an ANOVA (not shown) of the results of the analysis of the discriminant analysis results yielded a single large effect for DIST, which accounted for over 91% of the *total* variation in the HA-Rand index.

The data generated by the different combinations of R and ANGLE do not yield clustering problems which differ in their inherent difficulty. The results portrayed in Figure 6 are *not* an artifact of the method used to generate the data. Instead, they reflect a real aspect of the behavior of clustering methods.

The Effect of Within-group Correlation

The results in Figure 6 can best be understood according to the difference between the Euclidean (d_E) and Mahalanobis (D_M) measures of distance. The clustering methods used in this study are all based on (squared) Euclidean distances. Although Ward's method is often described as minimizing the "error-sum-of-squares," or the trace of the pooled within-group covariance matrix, these criteria are equivalent to operating on a function of the squared Euclidean distance (Wishart, 1969). The same is true of the EML clustering method.

The top of Figure 8 illustrates the difference between d_E and D_M. If $r_{xy} = .7$, the ellipse is the set of points which are $D_M = 2$ away from P. The set of points which are $d_E = 2$ away from P would form a circle. Only in the special case when the correlation is 0 do the two measures agree. This difference causes some inter-point distances to be ordered differently by the two measures; $D_M(PQ) = 3.24$ and $D_M(PR) = 1.63$, but $d_E(PQ) = 1.80$ and $d_E(PR) = 2.12$. The Mahalanobis distance is the appropriate representation, but clustering methods using d_E consider

Point Q closer than Point R, and so have a tendency to join Point Q with P, rather than joining Point R with Point P.

Insert Figure 8 about here

The bottom half of Figure 8 illustrates how the angle of the vector between the subgroup centroids affects the probability of a point like Q occurring. The left panel shows the $D_M = 2$ ellipses for two subgroups when the vector between the centroids forms an angle θ of 45° with the x-axis. D_M between the subgroup centroids is 4. The right panel shows the same ellipses when $\theta = 135°$. The two vectors in each figure represent distances between Point P, and Points Q and R. When $\theta = 135°$, it is much more likely than an observation like Point Q will be sampled from the other subgroup than it is when $\theta = 45°$. The tendency to join Point Q to Point P before joining Point R to Point P thus has the potential to more severely degrade recovery for subgroups configured as in the right panel than those configured as in the left panel. Thus, the discrepancy between the two distance measures can and does degrade subgroup recovery, as the results of Study 1 and Study 2 clearly show. Using d_E causes clustering methods to misrepresent the space.

Discussion

At this point we return to the questions posed at the beginning of this paper.

1) Does the within-group covariance structure effect the ability of commonly used clustering algorithms to recover known subgroup structure?

The answer to this question is a qualified yes. Strong effects were found for the within-group correlation. Within-group correlation which did not coincide with the direction of separation in subgroup centroids was associated with lower recovery for all clustering methods; within-group correlation which did coincide with the direction of separation was associated with higher recovery. However, this study only considered the bivariate case for two subgroups. It is not known to what extent this result is generalizable to higher dimensional problems or problems with more subgroups.

The relative size of the within-group variances was found to interact with the clustering algorithm used. Average linkage produced somewhat better recovery with homogeneous variances; in both studies average linkage was one of the better algorithms under these conditions. On the other hand, Ward's method, EML and the beta-flexible clustering method produced superior recovery when the larger subgroup was more compact, i.e., had a smaller variance. However, because of the limited nature of the manipulation of the within-group variances, these findings must be considered somewhat preliminary.

2) Do clustering methods differ in their sensitivity to within-group covariance structure?

All of the clustering methods examined were affected in the same manner by the within-group correlation. Moreover, the interaction of clustering method with within-group correlation was not identified as a salient effect in either Study 1 or Study 2.

The effect of within-group variances did differ across the clustering algorithms. As was noted above, the size of these effects was much smaller than those for within-group correlation. However, none of the clustering methods yielded recovery which was completely independent of the within-group variance, and the size of the difference in means between the best and worst conditions were relatively similar across methods.

3) Which clustering methods work best?

Comparison of the clustering methods is complicated by the salient interactions which were detected between Method and VAR in Study 1, and between Method and PROB in both Study 1 and Study 2. Ideally, one would choose the method which performed best for the levels of VAR and PROB which are most similar to the data at hand. However, in most clustering applications the analyst will not have this type of knowledge. Thus, it is of interest to compare the overall performance of the clustering methods.

Overall means for each clustering method are presented in the top of Table 6. The ranks are based on paired t-tests performed on all pairwise comparisons of clustering methods (using Shaffer's (1986) modification to the Bonferroni correction with familywise Type 1 error rate of $\alpha=0.05$ for each study). Methods which differ in rank correspond to statistically significant differences in recovery. The column labeled "Average" is simply the rank of the average of the two ranks. To help put the differences in Table 6 in perspective, the average change in the HA-Rand statistic of moving a single observation from one subgroup to another is approximately 0.003.[3]

The bottom half of Table 6 used Cliff's (1993) method of comparing the order of two distributions. The procedure estimates the probability that a randomly sampled observation from one distribution is higher than a randomly sampled observation from another distribution. Pairwise comparisons of clustering methods (with the Shaffer-Bonferroni correction) used a modified version of Cliff's (1992) program PAIRDEL1. This approach resulted in one of three decisions for each pair of clustering methods: a) Method A is higher (better recovery) than Method B; b) Method B is higher than Method A; or c) the methods do not significantly differ. Pairwise relations were then converted into ranks, based on the number of clustering methods which were significantly higher or lower than a given clustering method.

[3] This number was derived through brute force combinatorics. For each value of PROB, the number in each subgroup determined the fixed column totals of a 2 x 2 table. The cell counts were then systematically varied over the possible range, and the HA-Rand index computed for each combination. The difference in HA-Rand index was then computed for each move of one entity from one cluster (row) to the other, and weighted by the number of such moves. These weights were then used to compute the expected value of the difference for that value of PROB. F nally, the means of the two levels of PROB were averaged.

In general, the two approaches give similar results, although there are some striking differences because of the different hypotheses. For example, comparison of means in Study 1 indicates that EML gives the best recovery, while the ordinal procedure indicates that both Ward's method (using squared distances) and the beta-flexible method with $\beta=-.25$ outperform EML. For a number of data sets, Ward's method (for example) yielded cluster solutions with slightly higher ($|\text{diff}_{\text{HA-Rand}}| < .05$) recovery than EML. In a smaller number of cases, however, EML yielded substantially higher recovery than Ward's method, which causes EML to yield a higher mean HA-Rand index. Which clustering method is better depends upon the definition of better; Ward's method yields slightly better recovery in a number of cases, but occasionally EML gives much better recovery.

In spite of the differences, there is substantial agreement between the approaches as to which clustering methods give the best recovery for data of the type used in this study. Although the ordering differed slightly, EML, Ward's method (using squared distance), beta-flexible method with $\beta=-.25$, and average linkage (using distance) were identified as the four best methods by both approaches. The behavior of each of the methods is now briefly examined, and related to results from other studies of clustering methods.

Overall, SAS's EML method produced the highest mean recovery of subgroup structure, followed closely by Ward's method and the beta-flexible clustering algorithm with $\beta=-.25$. The EML method was devised (SAS Institute, 1988) to alleviate the tendency of Ward's method to produce equal-sized subgroups. The results in Figure 2 support this intention. The behavior of the EML method is less sensitive to the subgroup size than is Ward's method. EML was clearly superior to the beta-flexible and Ward's method for unequal probabilities of subgroup membership, and clearly inferior for equal probabilities. The EML method has not been included in previous comparisons of clustering methods, but its performance here argues that it should be included in future studies. On a practical note, however, the EML method is computationally intensive, and typically required 16-17 times as long to run as did any of the other algorithms.

The beta-flexible method produced good recovery, consistent with Milligan (1989a). Also, consistent with Milligan's findings, the value of $\beta = -.25$ produced the best overall value of recovery. However, Figure 2 reveals that the optimal value of β depended upon the relative size of the subgroups. For equal sized subgroups, recovery increased with smaller values of β, while the opposite was true when the subgroups differed in size. Although more work is needed to explore further the optimum value of β, Milligan's suggestion of $\beta = -.25$ is the best advice at present. While this paper was in preparation, a modified procedure was proposed which appears to be less sensitive to the value of β (Belbin, Faith, & Milligan, 1992). Recent work (Donoghue, 1994b) indicates that the modified procedure may yield superior recovery.

In Study 1, the average linkage method produced recovery which was substantially below that of the methods discussed above, while in Study 2, average linkage was second only to EML. This result stems from the superior recovery by average linkage for subgroups with homogeneous variances. Figure 2 reveals that average linkage is the <u>only</u> method which is insensitive to subgroup size, while Ward's method is very sensitive. In their review, Milligan and Cooper (1987) report that the average linkage method produces superior recovery to that of Ward's method for non-overlapping clusters, while Ward's method is superior for overlapping clusters, especially those based on mixtures of normal distributions. The results obtained here support the second half of this statement. Yet, when D_λ 6, the overlap between the subgroups is all but nonexistent. Figure 3 reveals that even in this condition, Ward's method still produced higher average recovery than did the average linkage method. Finally, using Euclidean distance produced slightly higher recovery than did using the squared distance.

Consistent with previous work (e.g., Milligan, & Cooper, 1987, and the references cited therein) complete linkage and single linkage produced cluster recovery which was inferior to that of most or all of the other methods examined. Unfortunately, these methods are widely available and continue to be used in applications, in spite of the large body of results testifying to the fact that better alternatives are available. The sharp schism between the "theoretical" and the "empirical" value of the single linkage algorithm is particularly bothersome. Based on theoretical considerations, such as its relationship to the minimum spanning tree of graph theory, several authors find single linkage to be the most attractive of the clustering methods, and continue to recommend it on those grounds (e.g., IMS Panel on Discrimination, Classification, and Clustering, 1989). Yet this method uniformly produced the lowest recovery of the methods examined. This result is routinely found in comparisons of the ability of clustering algorithms to recover subgroup structure (e.g., Milligan, & Cooper, 1987, and the references therein).

Finally, the pseudo-Ward algorithm produced somewhat higher recovery than did the true Ward's method for equal-sized subgroups, but performed much worse for unequal-sized subgroups. This finding resulted in lower overall recovery for the pseudo-Ward procedure. Coupled with the method's *ad hoc* nature, there is little to recommend the pseudo-Ward method.

Future Research: The Appropriate Choice of Similarity Measure

The choice of a similarity measure is not an easy one. Three commonly mentioned possibilities are: (a) Euclidean distance, (b) Mahalanobis distance, and (c) clustering based upon principal component scores. The problems with these obvious choices will be discussed, and then some possible alternatives will be discussed.

Three Obvious Choices

The Euclidean distance has clear drawbacks, as the results of Study 1 and Study 2 illustrate. As was discussed above, by ignoring the within-group correlations of variables, d_E can seriously degrade cluster recovery.

Ideally, one would cluster based on the Mahalanobis distance measure, D_M. In practice however, D_M is difficult to compute, because of the difficulty in estimating the appropriate Σ matrix. The correct choice is the pooled within-group covariance matrix, Σ_w:

$$\boldsymbol{\Sigma}_{W} = \sum_{g=1}^{G} \pi_g \boldsymbol{\Sigma}_g \quad .$$

Without knowing the subgroup structure, it is not possible to compute the Σ_g matrices. An alternative is to use the overall covariance matrix, Σ_T:

$$\boldsymbol{\Sigma}_{T} = \frac{1}{N}(\boldsymbol{x}-\overline{\boldsymbol{x}})'(\boldsymbol{x}-\overline{\boldsymbol{x}}) \quad .$$

However, a standard result from MANOVA gives:

$$\boldsymbol{\Sigma}_{T} = \boldsymbol{\Sigma}_{B} + \boldsymbol{\Sigma}_{W} \quad .$$

The inclusion of Σ_B can give Σ_T very different properties from Σ_W. Hartigan (1975, p. 63) provides an example demonstrating that D_M based on Σ_T can result in worse cluster recovery than using d_E.

A third alternative procedure is to cluster based upon principal component scores rather than the original variables. The chief advantage of this method is that principal component scores are mutually orthogonal; because d_E implicitly assumes orthogonal variables, clustering based on the principal components seems to remove the objections to this measure. The difficulty with principal components is identical to that of D_M. The ideal is to perform principal components analysis based on Σ_W. In practice, however, Σ_W is not known, and so component scores are computed based on

Σ_T (or the correlation matrix R_T). As is the case with D_M, the inclusion of Σ_B can give Σ_T very different properties from Σ_W. Chang (1983) gives a theoretical analysis of the difficulties involved and provides a clear example of the problems with this approach.

Alternative Procedures

At least three distance-based procedures have been suggested to improve measures of similarity for clustering: multidimensional scaling; Art, Gnanandesikan, and Kettenring's (1982) method to estimate the pooled within-group covariance matrix; and De Soete's (1986, 1988) algorithm to weight variables so as to maximize the agreement with ultrametric inequality. Each of these procedures will now be briefly described. No attempt is made to examine non-distance related measures of similarity, such as q-correlations. Interested readers are referred to standard sources on clustering, such as Everitt (1974) and Lorr (1983) for a discussion of non-distance similarity indices.

Multidimensional scaling (MDS) has a long history in behavioral sciences as a tool to investigate the relationships among entities. Historically, MDS has most commonly been used to obtain low dimensional representation of the interrelationships of entities (e.g., stimuli) in order to describe the dimensions (or the entire space). However, there have been applications of MDS to clustering as well (e.g., Boch, 1986; DeSarbo, Carroll, Clark, & Green, 1984; DeSarbo, Howard, & Jedidi, 1991; De Soete, DeSarbo, & Carroll, 1985). A variety of MDS software is available, and although I am aware of no studies specifically examining the use of MDS in relation to within-group covariance structure, it is surely an area which bears exploration.

Art, Gnanandesikan, and Kettenring (1982) devised a method to estimate the pooled within-group covariance matrix without knowledge of the subgroup structure. Such an estimate would allow one to compute a valid D_M measure, and in theory could greatly improve the accuracy of cluster recovery. A modified version of this estimate is computed by SAS's PROC ACECLUS (SAS Institute, 1988), including the option to output a matrix of inter-entity similarities for further analysis. To date, I am not aware of any published studies of ACECLUS, although preliminary work by Donoghue (1994b) indicates that the method is promising. Further studies would be useful, especially because the procedure requires the user to specify the values of certain parameters. Guidance in the use of these parameters would certainly help practitioners in applying the method.

De Soete (1986, 1988) developed an algorithm which determines weights to apply to variables in computing a distance measure. The algorithm is based upon a relation known as the "ultrametric inequality," which states that for any three points *i, j, and k,* the distances between the points should satisfy the relation:

$$d_{ij} \leq \max(d_{ik}, d_{jk}) \quad .$$

This is equivalent to requiring all sets of three points to lie on an acute isosceles (or equilateral) triangle. Johnson (1967) and Milligan (1979) demonstrated the relationship between the ultrametric inequality and many commonly used hierarchical clustering algorithms, and Milligan and Isaac (1980) give simulation results which provide support for the utility of the conceptualization. Two studies (Milligan, 1989b, Donoghue, 1994a) have examined De Soete's algorithm, and found that it greatly improved cluster recovery when the data contained one, two, or three irrelevant dimensions. Further studies to evaluate this method would clearly be useful.

Conclusion

This study was intended to constitute a first step in understanding the sensitivity of clustering algorithms to within-group covariance structures. Clustering algorithms *were* found to be affected by the within-group covariance structure. Moderate effects were found for the interaction of clustering method with within-group variance. All of the clustering methods examined were strongly affected by the within-group correlation and its relationship to the vector which separated the centroids.

The results for within-group correlation were interpreted in terms of the similarity measure used in this study (Euclidean distance), and possible alternative measures were identified for future research. But these results also point out a larger issue. In the earlier clustering literature, the properties of measures of similarity were carefully considered and debated. One well-known interchange involved the Euclidean distance. Cronbach and Gleser (1953) examined the measure and strongly supported its use. Overall (1964) disagreed, pointing out that the formula for d_E relies on orthogonal variables. He advocated the use of D_M. Heerman (1965) took a different view, contending that nonorthogonality was irrelevant, and supporting the use of d_E. In the past decade, however, this debate has largely disappeared from the clustering literature. The vast majority of recent studies comparing clustering methods have used orthogonal variables and Euclidean distance as a measure of similarity. Although these studies have told much about certain aspects of the process of cluster analysis, they have tended to downplay the importance of the measure of similarity. While no single study can fully examine each of the facets of cluster analysis, it is important that our simulations be varied enough to reflect adequately the richness of the empirical data which researchers are likely to encounter in clustering applications.

References

Anderson, J. C., & Gerbing, D. W. (1984). The effect of sampling error on convergence, improper solutions, and goodness-of-fit indices for maximum likelihood confirmatory factor analysis. Psychometrika, 49, 155-173.

Art, D., Gnanandesikan, R., & Kettenring, R. (1982). Data-based metrics for cluster analysis. Utilitas Mathematica, 21A, 75-99.

Belbin, L., Faith, D. P., & Milligan, G. W. (1992). A comparison of two approaches to beta-flexible clustering. Multivariate Behavioral Research, 27, 417-433.

Blashfield, R. K. (1976). Mixture model tests of cluster analysis: Accuracy of four agglomerative hierarchical methods. Psychological Bulletin, 83, 377-388.

Blashfield, R. K., & Morey, L. C. (1980). A comparison of four clustering methods using MMPI Monte Carlo data. Applied Psychological Measurement, 4, 57-64.

Boch, H.-H. (1986). Multidimensional scaling in the framework of cluster analysis. In P. O. Dregens, H.-J. Hermes, & O. Opitz (Eds.) Sudien zur Klassifikation, Bd. 17 (SK 17), pp. 247-258. Frankfurt/Main, Germany: Indeks Verlag.

Breckenridge, J. N. (1989). Replicating cluster analysis: Method, consistency, and validity. Multivariate Behavioral Research, 24, 147-161.

Chang, W-C. (1983). On using principal components before separating a mixture of two multivariate normal distributions. Applied Statistics, 32, 267-275.

Cronbach, L. J., & Gleser, G. C. (1953). Assessing similarity between profiles. Psychological Bulletin, 50, 456-473.

Cliff N. (1992). PAIRDEL1.BAS: Program for computing matched-data d-statistics [computer program]. Los Angeles: Psychology Department, University of Southern California.

Cliff, N. (1993). Dominance relations: Ordinal analyses to answer ordinal questions. Psychological Bulletin, 114, 494-509.

DeSarbo, W.S., Carroll, J. D., Clark, L. A., & Green, P. E. (1984). Synthesized clustering: A method for amalgamating alternative clustering bases with differential weighting of variables. Psychometrika, 49, 57-78.

DeSarbo, W., Howard, D. J., & Jedidi, K. (1991). MULTICLUS: A new method for simultaneously performing multidimensional scaling and cluster analysis. Psychometrika, 56, 121-136.

De Soete, G., DeSarbo, W. S., & Carroll, J. D. (1985). Optimal variable weighting for hierarchical clustering: An alternating least squares algorithm. Journal of Classification, 2, 173-192.

De Soete, G. (1986). Optimal variable weighting for ultrametric and additive tree clustering. Quality and Quantity, 20, 169-180.

De Soete, G. (1988). Software abstract - OVWTRE: A program for optimal variable weighting for ultrametric and additive tree fitting. Journal of Classification, 5, 101-104.

Donoghue, J. R. (1994a). Variable screening for cluster analysis. Research Report No. RR-94-36. Princeton, NJ: Educational Testing Service.

Donoghue, J. R. (1994b, April). Comparing the effectiveness of cluster analysis weighting procedures for within-group covariance structure: The Bivariate Case. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Dreger, R. M., Fuller, J., & Lemoine, R. L. (1988). Cluster analysis of seven data sets by means of some or all of seven clustering methods. Multivariate Behavioral Research, 23, 203-230.

Everitt, B. S. (1974). Cluster analysis. London: Halstead Press.

Fisher, R. K. (1936). The use of multiple measurements in taxonomic problems. Annals of Eugenics, 7, 179-188.

Gerbing, D. W., & Anderson, J. C. (1985). The effects of sampling error and model characteristics on parameter estimation for maximum likelihood confirmatory factor analysis. Multivariate Behavioral Research, 20, 255-271.

Hartigan, J. A. (1975). Clustering algorithms. New York: John Wiley & Sons.

Heerman, E. F. (1965). Comments on Overall's "Multivariate methods for profile analysis." Psychological Bulletin, 63, 128.

Hubert, L., & Arabie, P. (1985). Comparing partitions. Journal of Classification, 2, 193-218.

IMS Panel on Discriminant Analysis, Classification, and Clustering (1989). Discriminant analysis and clustering. Statistic Science, 4, 34-69.

Johnson, S. C. (1967). Hierarchical clustering schemes. Psychometrika, 32, 47-62.

Lance, G. N., & Williams, W. T. (1967). A general theory of classificatory sorting strategies: 1. Hierarchical systems. Computer Journal, 9, 373-380.

Lorr, M. (1983). Cluster analysis for social scientists. San Francisco: Jossey-Bass Publishers.

McLachlan, G. J., & Basford, K. E. (1988). Mixture models: Inference and applications to clustering. New York: Marcel Dekker.

Mezzich, J. E., & Solomon, H. (1980). Taxonomy and behavioral science. New York: Academic Press.

Milligan, G. W. (1979). Ultrametric hierarchical clustering algorithms. Psychometrika, 44, 343-346.

Milligan, G. W. (1980). An examination of the effects of six types of error perturbation on fifteen clustering algorithms. Psychometrika, 45, 325-342.

Milligan, G. W. (1981). A Monte Carlo study of thirty internal criterion measures for cluster analysis. Psychometrika, 46, 187-199.

Milligan, G. W. (1985). An algorithm for generating artificial test clusters. Psychometrika, 50, 123-127.

Milligan, G. W. (1989a). A study of the beta-flexible clustering method. Multivariate Behavioral Research, 24, 163-176.

Milligan, G. W. (1989b). A validation study of variable weighting algorithm for cluster analysis. Journal of Classification, 6, 53-71.

Milligan, G. W., & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. Psychometrika, 50, 159-170.

Milligan, G. W., & Cooper, M. C. (1986). A study of comparability of external criteria for hierarchical cluster analysis. Multivariate Behavioral Research, 21, 441-458.

Milligan, G. W., & Cooper, M. C. (1987). Methodology review: Clustering methods. Applied Psychological Measurement, 11, 329-354.

Milligan, G. W., & Cooper, M. C. (1988). A study of variable standardization. Journal of Classification, 5, 181-204.

Milligan, G. W., & Isaac, P. D. (1980). The validation of four ultrametric clustering algorithms. Pattern Recognition, 12, 41-50.

Overall, J. E. (1964). Note on multivariate methods for profile analysis. Psychological Bulletin, 61, 195-198.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association, 66, 846-850.

SAS Institute, Inc. (1988). SAS/STAT user's guide: Release 6.03 edition. Cary, NC: Author.

Scheibler, W., & Schneider, D. (1985). Monte Carlo tests of the accuracy of cluster analysis algorithms: A comparison of hierarchical and nonhierarchical methods. Multivariate Behavioral Research, 20, 283-304.

Shaffer, J. P. (1986). Modified sequentially rejective multiple test procedures. Journal of the American Statistical Association, 81, 826-831.

Siotani, M. (1982). Large sample approximations and asymptotic expansions of classification statistics. In P. R. Krishnaiah and L. N. Kanal (Eds.), Handbook of statistics, Volume 2. pp. 61-100. New York: North-Holland Publishing Company.

Tabachnick, B. G., & Fidell, L. S. (1983). Using multivariate statistics. New York: Harper and Row.

Titterington, D. M., Smith, A. F. M., & Makov, U. E. (1985). Statistical analysis of finite mixture distributions. New York: John Wiley & Sons.

Ward, J. H. (1963). Hierarchical subgrouping to optimize an objective function. Journal of the American Statistical Association, 58, 236-244.

Wishart, D. (1969). An algorithm for hierarchical classifications. Biometrics, 25, 165-170.

Table 1

Descriptive Statistics for Observed D_M

Study 1

D_M	Mean	Std.	Min.	Q_1	Med.	Q_3	Max.
2	2.081	0.432	0.198	1.848	2.048	2.269	6.027
4	4.110	0.565	0.486	3.778	4.062	4.383	9.676
6	6.141	0.714	3.158	5.704	6.069	6.505	15.670

Table 2

Main Effects and Salient Interactions ($\eta^2_{alt} > .03$) for ANOVA of HA-Rand Index

Study 1

EFFECT	DF	SS	η^2_{alt}
M	9.	2052.64	.155
D	2.	20345.79	.645
P	1.	1244.26	.100
V	4.	157.86	.014
R1	6.	4750.52	.298
R2	6.	859.05	.071
M*D	18.	469.25	.040
M*P	9.	1819.41	.140
M*V	36.	349.73	.030
D*R1	12.	845.73	.070
P*R1	6.	695.19	.058
V*R1	24.	1044.81	.085
V*R2	24.	438.28	.038
R1*R2	36.	389.13	.034
D*P*R1	12.	460.42	.040
Error	279300.	11194.93	---

Abbreviations: M - clustering method, D - Mahalanobis distance between subgroup centroids, P - probability of subgroup membership, V - ratio of subgroup variances, R1 - correlation in larger subgroup (when $p_1 \neq p_2$), R2 - correlation in smaller subgroup (when $p_1 \neq p_2$)

Table 3

Mean Values of HA-Rand Index for

Correlation within Subgroup 1 (R1) by Variance Ratio Interaction

Study 1

R1	Ratio of Variances (Subgroup 1: Subgroup 2)				
	1:9	1:4	1:1	4:1	9:1
-0.9	.497	.432	.295	.172	.137
-0.7	.529	.484	.395	.333	.300
-0.3	.584	.572	.570	.574	.573
0.0	.615	.605	.623	.634	.646
0.3	.623	.625	.652	.661	.667
0.7	.621	.630	.655	.657	.679
0.9	.621	.627	.646	.658	.668

Table 4

Mean Values of HA-Rand Index for

Correlation within Subgroup 2 (R2) by Variance Ratio Interaction

Study 1

R2	Ratio of Variances (Subgroup 1: Subgroup 2)				
	1:9	1:4	1:1	4:1	9:1
-0.9	.376	.398	.466	.493	.510
-0.7	.480	.477	.499	.510	.512
-0.3	.584	.566	.536	.523	.521
0.0	.630	.601	.565	.528	.526
0.3	.653	.623	.583	.534	.527
0.7	.682	.561	.594	.548	.535
0.9	.684	.659	.592	.551	.538

Table 5

Main Effects and Salient Interactions ($\eta^2_{alt} > .03$) for ANOVA of HA-Rand Index

Study 2

Effect	DF	SS	η^2_{alt}
D	2.	3794.733	.706
P	1.	284.340	.152
M	9.	310.716	.164
A	7.	15.226	.010
R1	2.	131.960	.077
R2	2.	5.715	.004
D*M	18.	147.821	.085
P*M	9.	297.683	.158
P*R1	2.	64.150	.039
A*R1	14.	248.968	.136
R1*R2	4.	63.622	.039
P*A*R1	14.	58.258	.036
Error	38880	1581.488	---

Abbreviations: M - clustering method, D - Mahalanobis distance between subgroup centroids, P - probability of subgroup membership, A - angle θ between x-axis and vector separating subgroup centroids, R1 - correlation in larger subgroup (when $p_1 \neq p_2$), R2 - correlation in smaller subgroup (when $p_1 \neq p_2$)

Table 6

Rankings of Clustering Methods Based on Means

Method	Study 1		Study 2		Rank		
	Mean	Std.	Mean	Std.	Study 1	Study 2	Avg.
EML	.623	.402	.638	.395	1	1	1
Ward's Method, d^2	.619	.405	.587	.407	2	2	2
Beta-flexible, $\beta=-.25$	.616	.401	.565	.404	3	5	3
Average Linkage, d	.530	.432	.593	.421	7	2	4
Beta-flexible, $\beta=-.50$	.607	.407	.530	.411	4	6	5
Average Linkage, d^2	.528	.430	.593	.418	8	2	5
Ward's Method, d	.594	.420	.536	.419	5	6	7
Beta-flexible, $\beta=-.75$	.568	.418	.493	.412	6	9	8
Complete Linkage, d	.476	.405	.517	.416	9	8	9
Single Linkage, d	.342	.433	.313	.423	10	10	10

Rankings of Clustering Methods Based on Ordinal Comparisons

Methods	Study 1	Study 2	Avg.
Ward's Method, d^2	1[a]	2[c]	1
EML	4[b]	1	2
Beta-flexible, $\beta=-.25$	1[a]	4[d]	2
Average linkage, d	7	3[cd]	4
Beta-flexible, $\beta=-.50$	3[ab]	7	4
Ward's Method, d	5	6	6
Average linkage, d^2	7	4[d]	6
Beta-flexible, $\beta=-.75$	6	8	8
Complete linkage, d	9	8	9
Single linkage, d	10	10	10

[abcd] Methods with common superscripts did not significar differ from one another. These methods received different ranks due to differen relations with other methods.

Figure Captions

Figure 1. Study 1: Response surface of HA-Rand Index as a function of correlation within subgroup 1 (R1) and correlation within subgroup 2 (R2).

Figure 2. Study 1: Interaction of clustering method with probability of subgroup membership.

Figure 3. Study 1: Interaction of clustering method with Mahalanobis Distance between subgroup centroids.

Figure 4. Study 1: Interaction of clustering method with ratio of within-group variances.

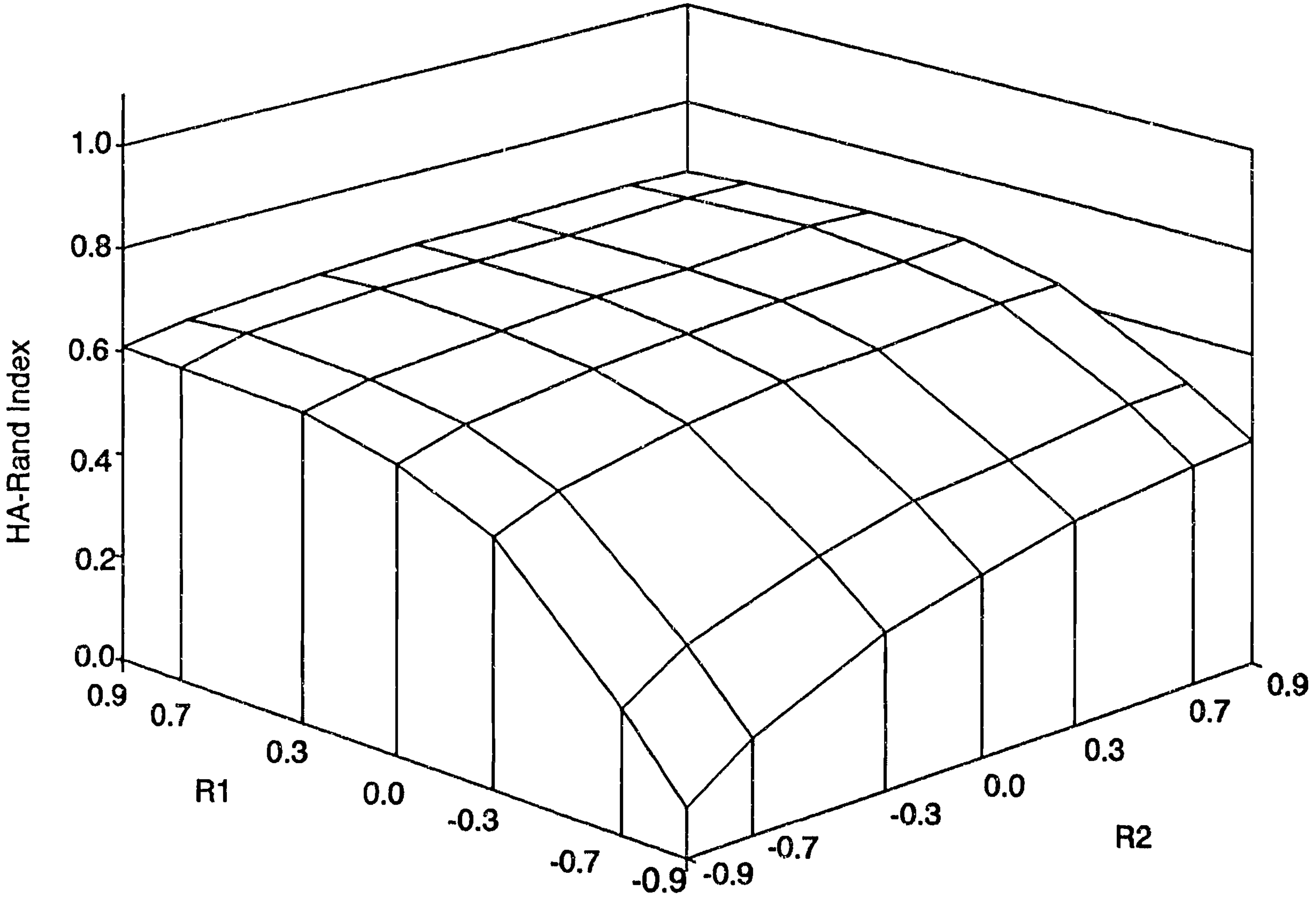
Figure 5. Study 1: Three-way interaction of probability of subgroups with Mahalanobis Distance between subgroup centroids with correlation within subgroup 1 (R1).

Figure 6. Study 2: Three-way interaction of probability of subgroups with correlation within subgroup 1 (R1) with angle θ formed between x-axis and vector between subgroup centroids.

Figure 7. Plot of centroid of subgroup 2 at various angles θ and within-group correlation used in Study 2: for conditions of equal within-group covariance matrices and Mahalanobis distance $D_M = 4$.

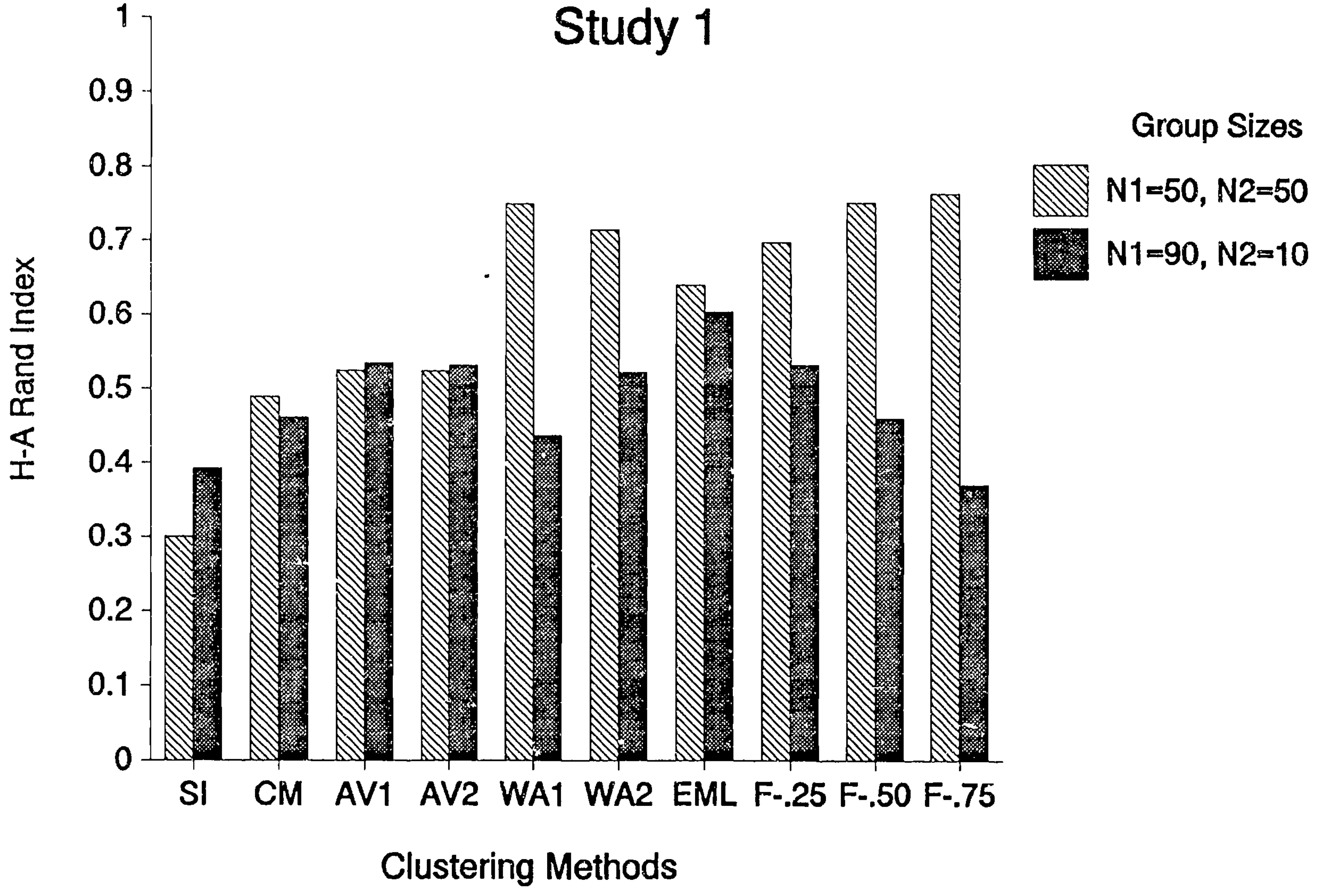
Figure 8. Plot of ellipse of equal Mahalanobis distance $D_M = 2$ around Point P, with two other points, Q and R. Plot of ellipse of equal Mahalanobis distance $D_M = 2$ around Point P, including additional subgroup with centroid at Mahalanobis distance $D_M = 4$ from P, where $\theta = 45°$ and $\theta = 135°$.

R1 by R2 Interaction



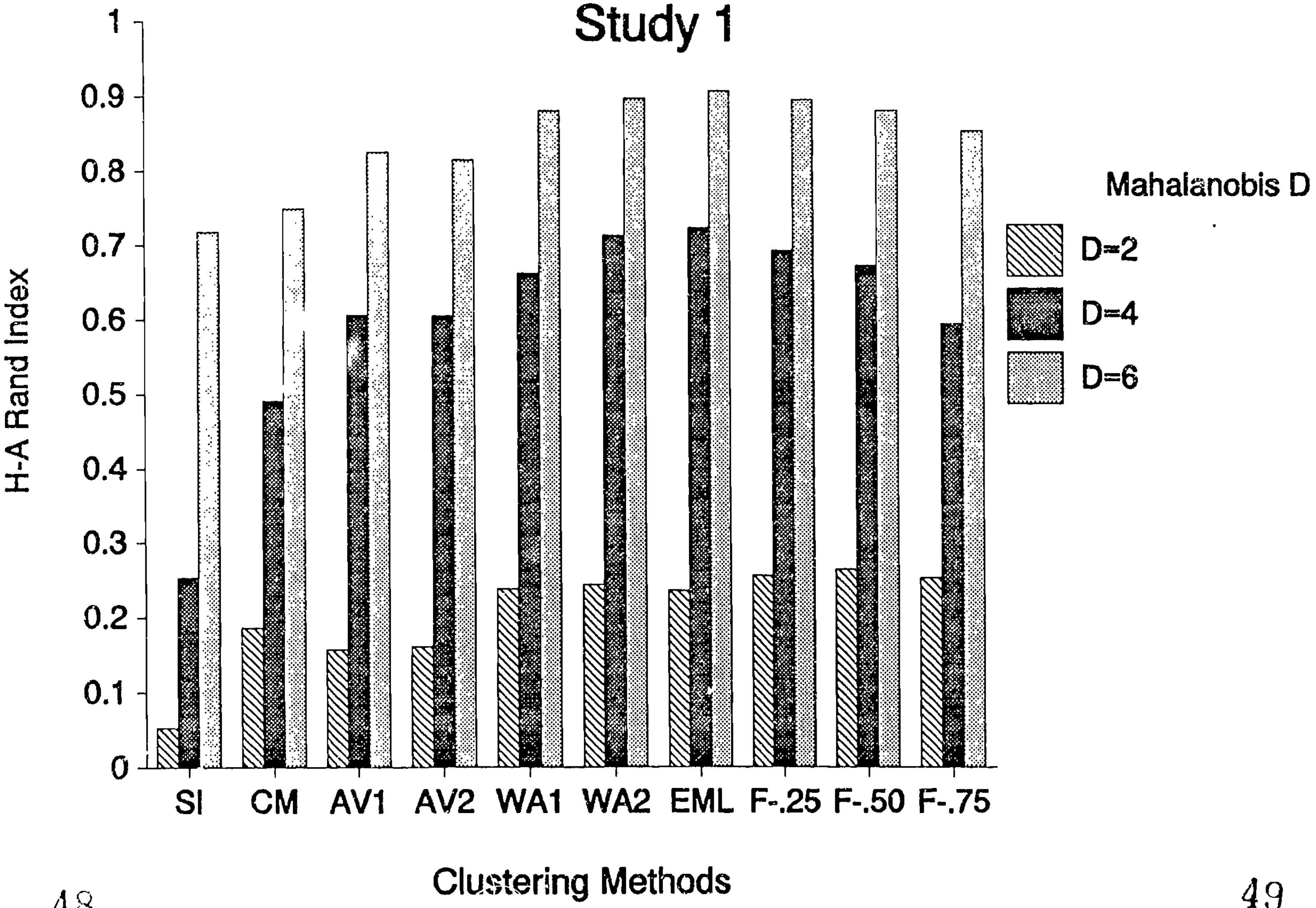
Method by Prob. Interaction

Study 1



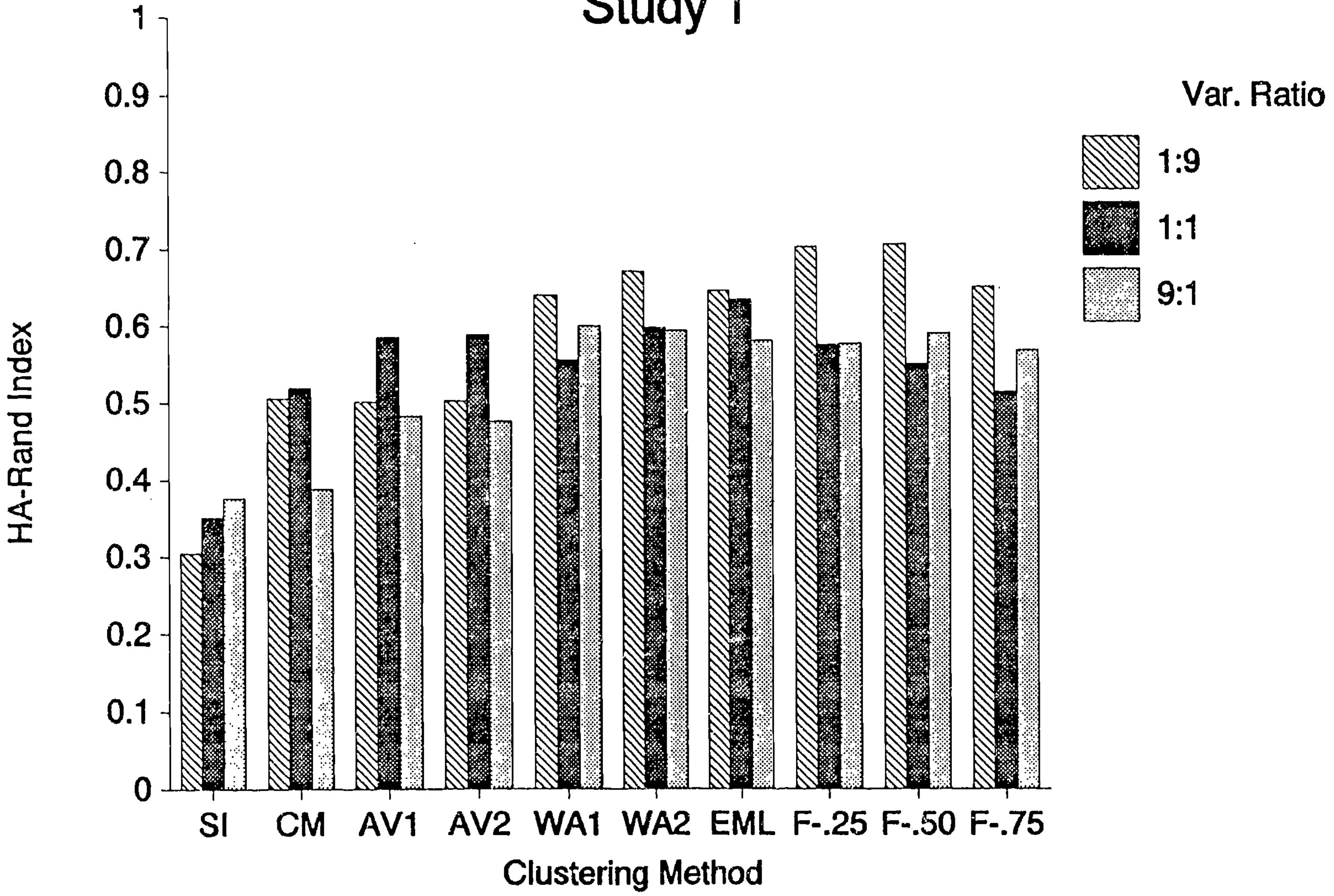
Method by Distance Interaction

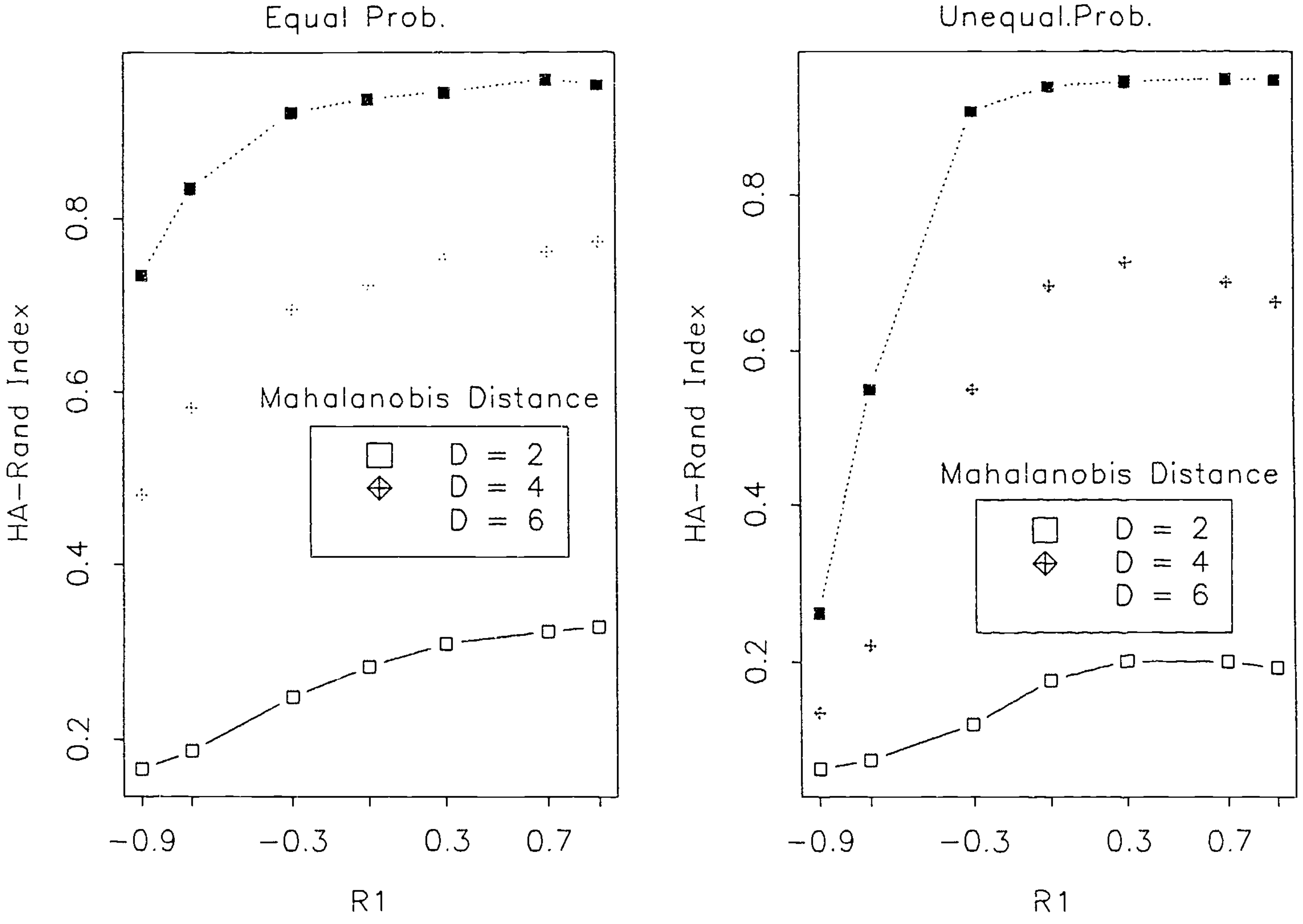
Study 1

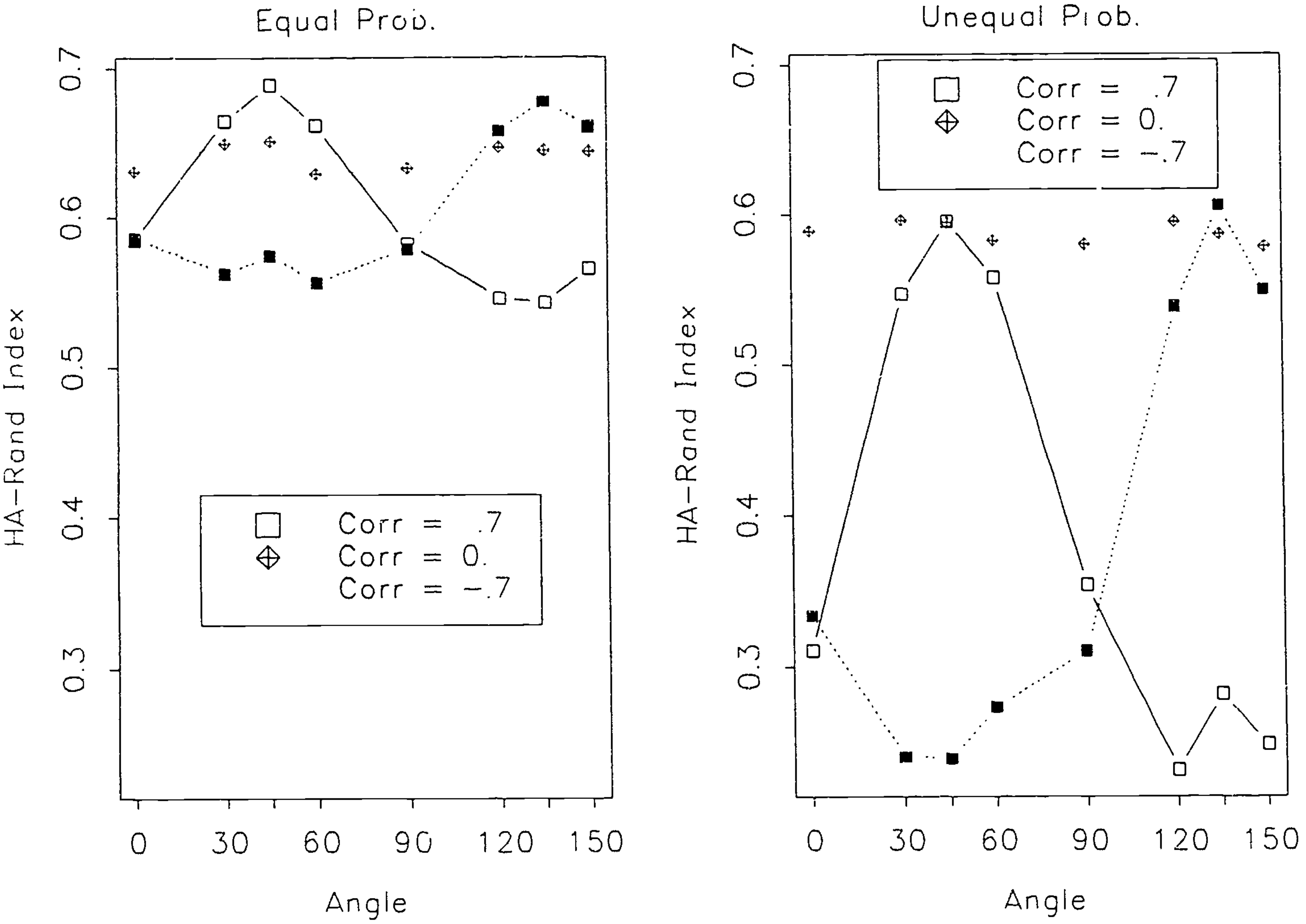


Method by COV Interaction

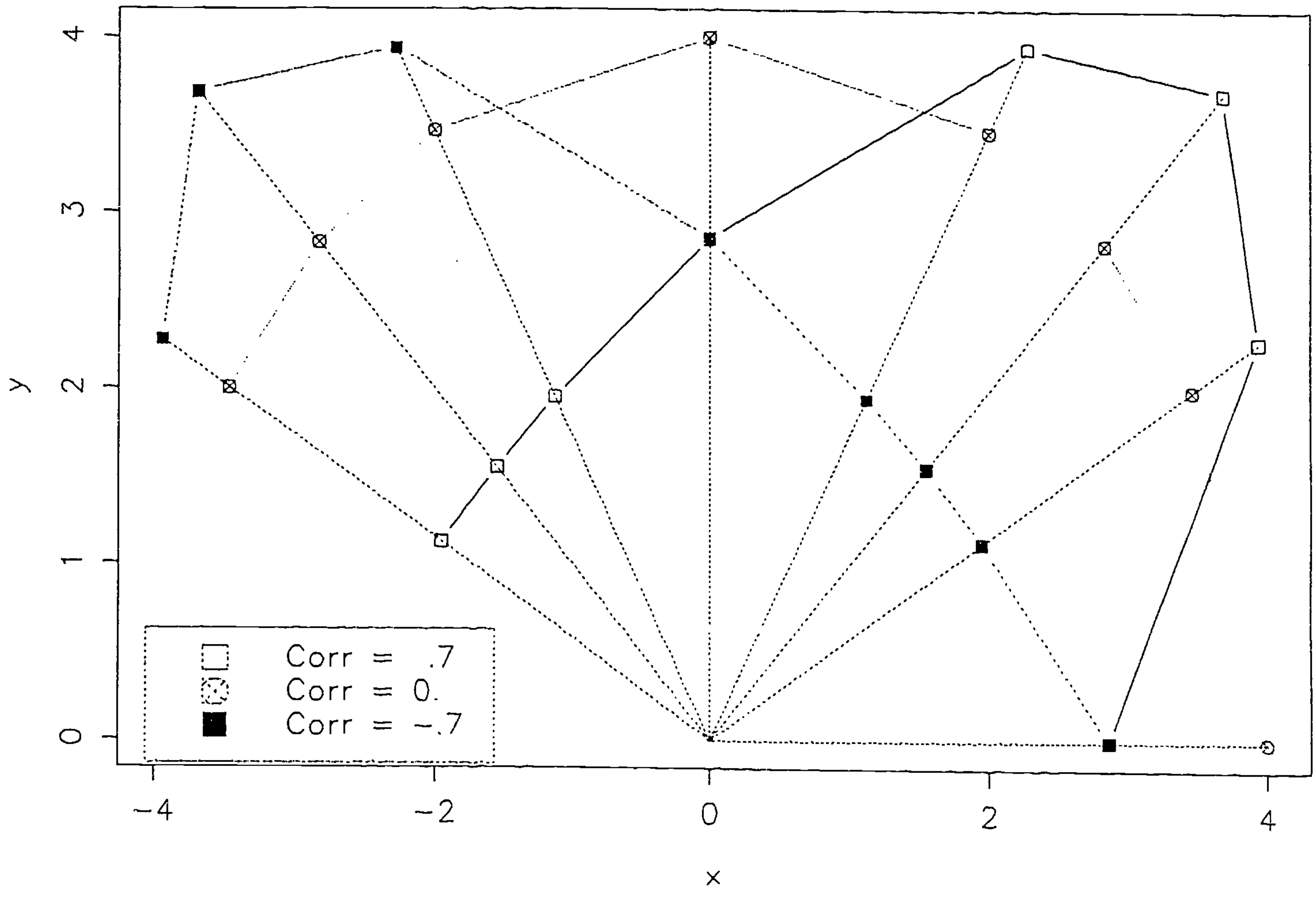
Study 1

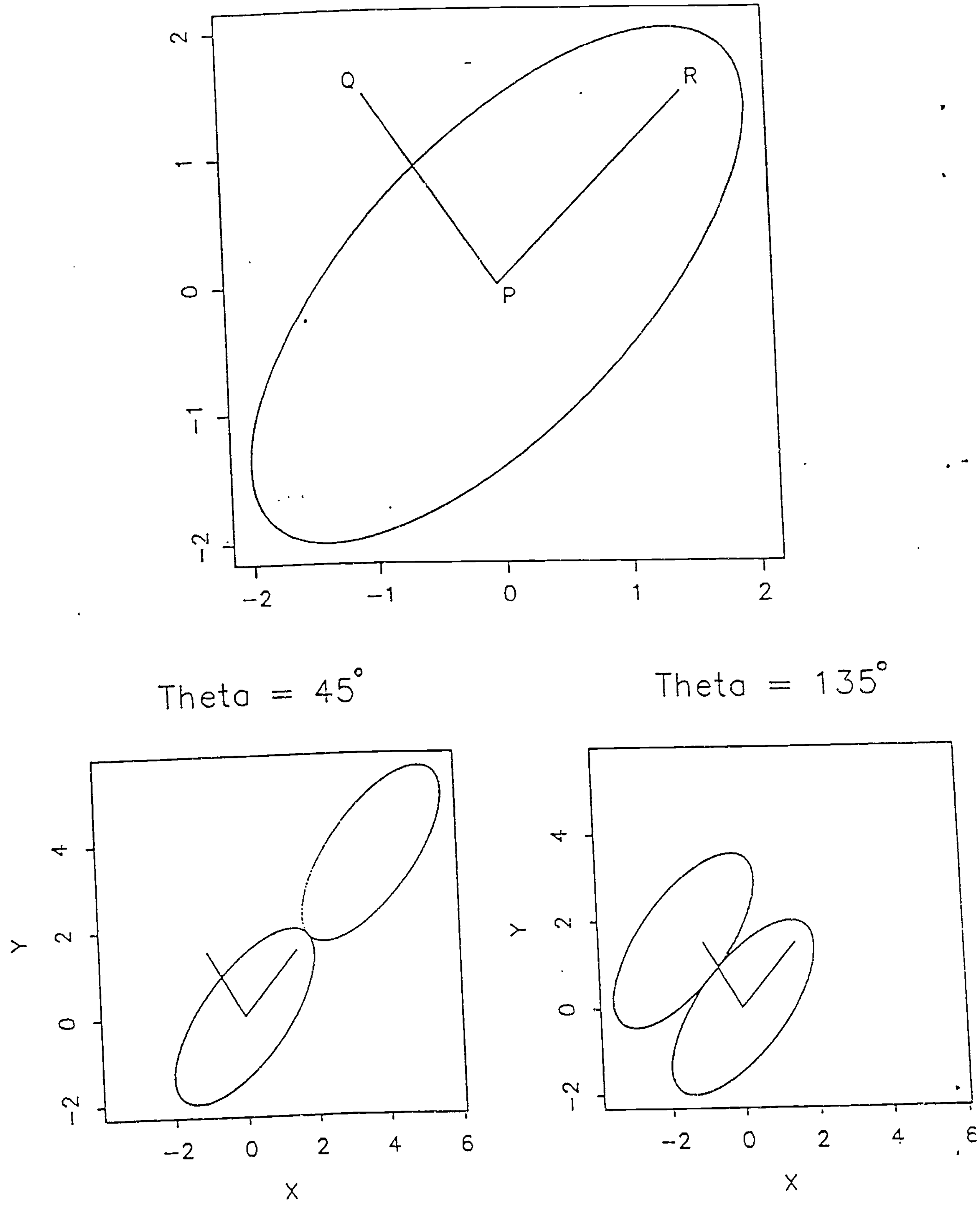



Equal Prob.
Unequal.Prob.
HA-Rand Index
0.2
0.4
0.6
0.8
Mahalanobis Distance
D = 2
D = 4
D = 6
−0.9
−0.3
0.3
0.7
R1


Equal Prob.
0.7
0.6
0.5
0.4
0.3
HA-Rand Index
□ Corr = .7
◇ Corr = 0.
Corr = -.7
0
30
60
90
120
150
Angle
Unequal Prob.
0.7
0.6
0.5
0.4
0.3
HA-Rand Index
□ Corr = .7
◇ Corr = 0.
Corr = -.7
0
30
60
90
120
150
Angle

Centroids by Corr.




Theta = 45°

Theta = 135°